# Uber vs Lyft Price Prediction

## Milestone: Project Report

Unnati Mehta

mehta.un@northeastern.edu

## 1. Problem Setting:

Which ride should you book, Uber or Lyft?

It is often a dilemma to choose between Uber or Lyft for which minimizes the cost. Although these types of services offered through apps may appear to be similar, there are distinctions between the two most prominent transportation network services in the United States.

- Adaptive pricing is a pricing strategy that involves adjusting the cost of a product or service based on various factors such as location and time of day. This means that the prices of a product or service may fluctuate and that the best way to compare the prices of different companies is by checking them in real-time using both apps on your phone.
- Knowing the most frequently used locations for pick-up and drop-off in Boston is important.

## 2. Problem Definition:

Carry out data gathering and preparation, exploration, and representation. Choose and apply machine learning techniques and evaluate their effectiveness.

Motive of the project:

- Price analysis for distances, time of the day, weather conditions, etc.
- Best ride category of the cab, Location of drop-off and pickup analysis
- Considering the significant external - predictors like:
    - Day of the Week
    - Climatic Conditions
    - Office / Non-office hours

## 3. Data Source:

The dataset of Lyft and Uber ride is from Kaggle Click here for Link: **Uber_Lyft_Dataset**

## 4. Data Description:

The data is for time period 11/2018 to 12/2018. The Boston Uber Lyft dataset likely includes several variables, or columns of data, related to each ride provided by the services. It comprises of 693072 data points with 56 out of which 10 are categorical and 46 are numerical variables and 1 target variable (Target: Price). This dataset includes information such as the location of pick-up and drop-off points, the time and date of rides, the fare charged, and the type of vehicle used.

## 5. Data Exploration:

- The dataset has a total of 693071 bookings for Uber and Lyft respectively with 57 attributes. For each car type the count is:

| | cab_type |
|---|---|
| **Uber** | 385663 |
| **Lyft** | 307408 |

- The local Taxi booking that is supported by Uber has been removed. The dimension of the dataset reduces as 637976 observations with 57 attributes. The count of only Uber reduces:

| | cab_type |
|---|---|
| **Uber** | 330568 |
| **Lyft** | 307408 |

- The dataset is checked for null values:

```
id                              0
timestamp                       0
hour                            0
day                             0
month                           0
datetime                        0
timezone                        0
source                          0
destination                     0
cab_type                        0
product_id                      0
name                            0
price                           0
distance                        0
surge_multiplier                0
latitude                        0
longitude                       0
temperature                     0
apparentTemperature             0
short_summary                   0
long_summary                    0
precipIntensity                 0
precipProbability               0
humidity                        0
windSpeed                       0
windGust                        0
windGustTime                    0
visibility                      0
temperatureHigh                 0
temperatureHighTime             0
temperatureLow                  0
temperatureLowTime              0
apparentTemperatureHigh         0
apparentTemperatureHighTime     0
apparentTemperatureLow          0
apparentTemperatureLowTime      0
icon                            0
dewPoint                        0
pressure                        0
windBearing                     0
cloudCover                      0
uvIndex                         0
visibility.1                    0
ozone                           0
```

```
sunriseTime                 0
sunsetTime                  0
moonPhase                   0
precipIntensityMax          0
uvIndexTime                 0
temperatureMin              0
temperatureMinTime          0
temperatureMax              0
temperatureMaxTime          0
apparentTemperatureMin      0
apparentTemperatureMinTime  0
apparentTemperatureMax      0
apparentTemperatureMaxTime  0
dtype: int64
```

- *This shows that there are no null values in the dataset.*

## Data Visualization:

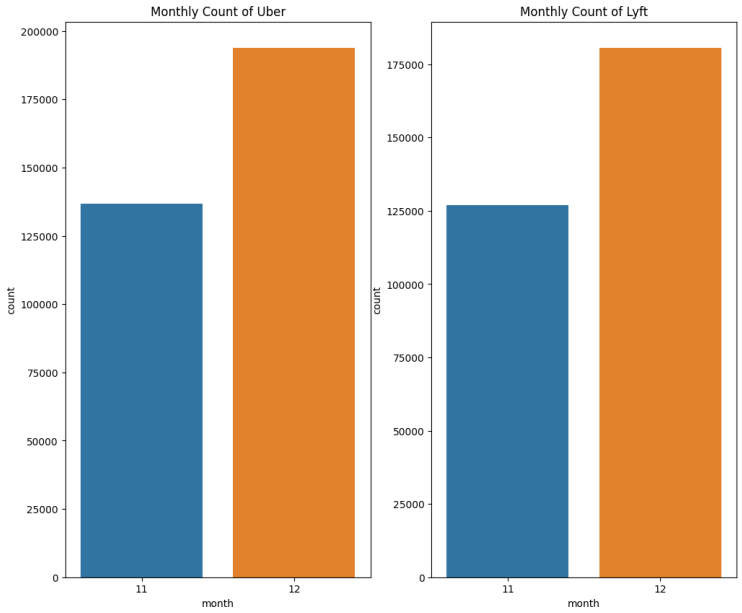



Figure 1: Monthly Count

- *From the above count plot, we observe the following:*

1. *The booking records are the months of November and December*
2. *We have 136846 records for November for Uber*
3. *We have 193722 records for December for Uber*
4. *We have 126925 records for November for Lyft*
5. *We have 180483 records for December for Lyft*

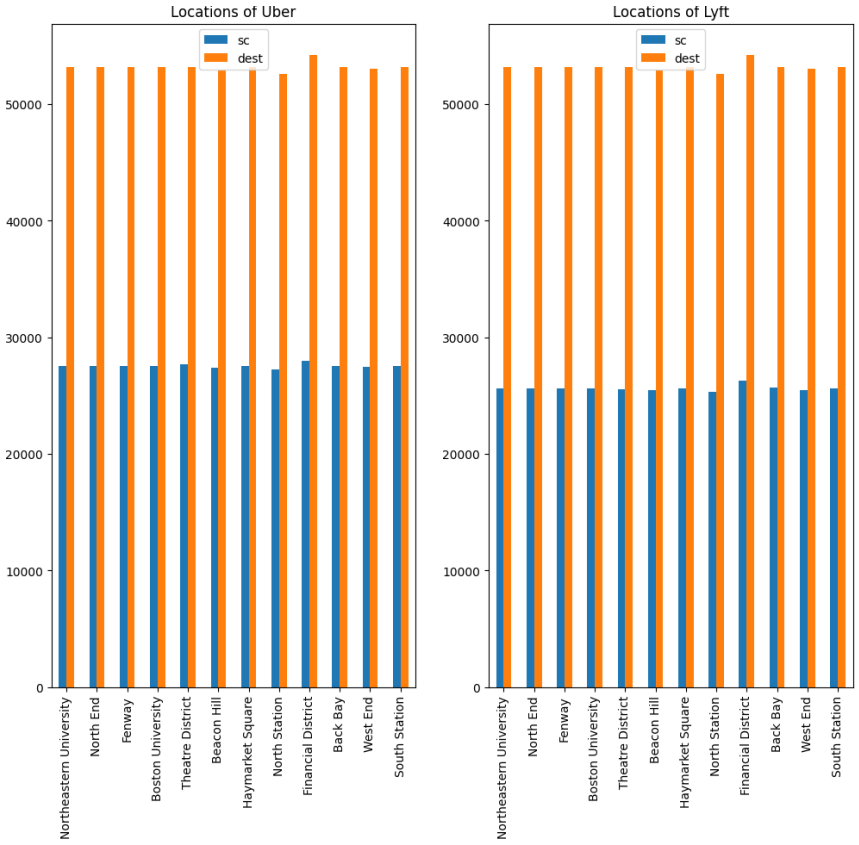



Figure 2: Location of Cabs booked.

- *From the above side-by-side bar chart, we observe that:*
- *FOR UBER:*

1. *Maximum cabs were booked from Back Bay (27546 cabs)*
2. *Minimum cabs were booked from West End (27492 cabs)*
3. *Maximum cabs were booked to Financial District (27954 cabs)*
4. *Minimum cabs were booked to North Station (27251 cabs)*

- *FOR LYFT:*

1. *Maximum cabs were booked from Financial District (26237 cabs)*
2. *Minimum cabs were booked from North Station (25326 cabs)*
3. *Maximum cabs were booked to Financial District (26238 cabs)*
4. *Minimum cabs were booked to North Station (25326 cabs)*

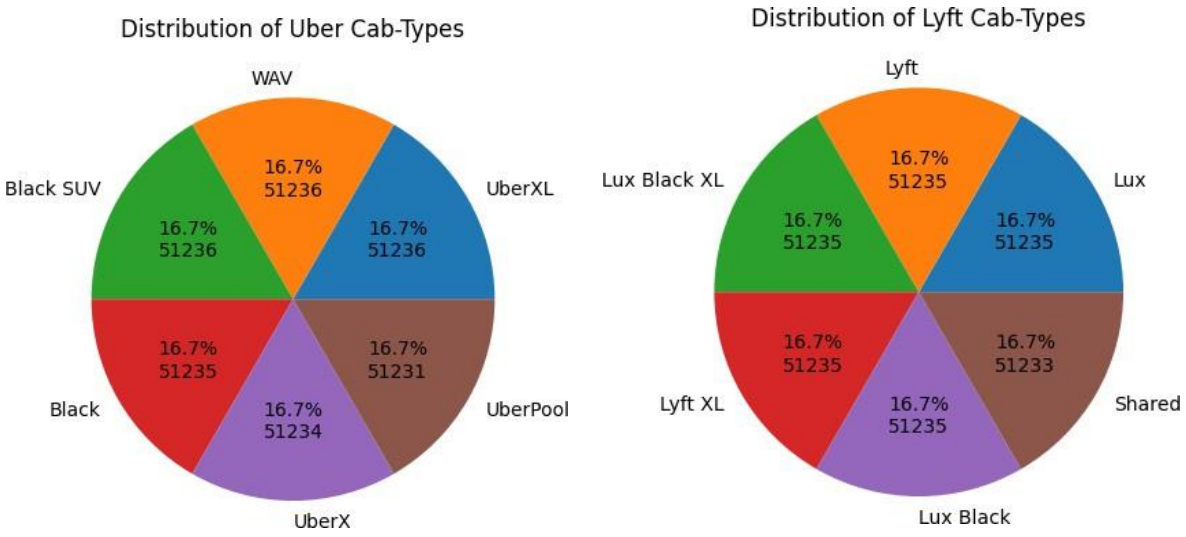



Figure 3: Distribution of Cab types

- *From the above pie-chart, we observe that:*

1. *An equal proportion of all the different types of cabs were booked.*

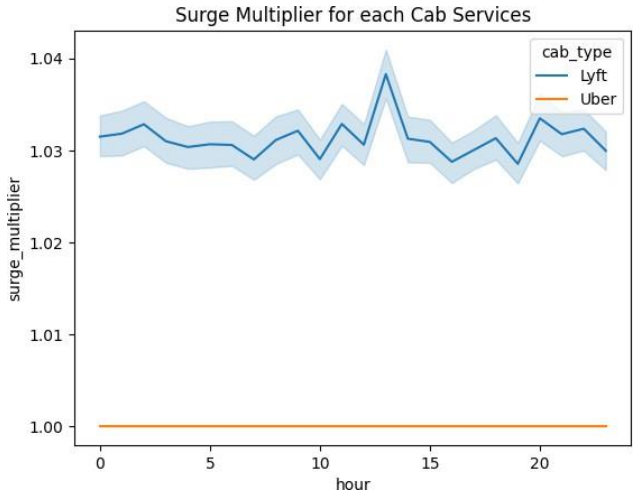



Figure 4: Surge Multiplier

- *From the above lineplot, we observe that:*

1. *The surge multiplier remains the same throughout the day for Uber.*
2. *The surge multiplier is the highest in the afternoon for Lyft.*

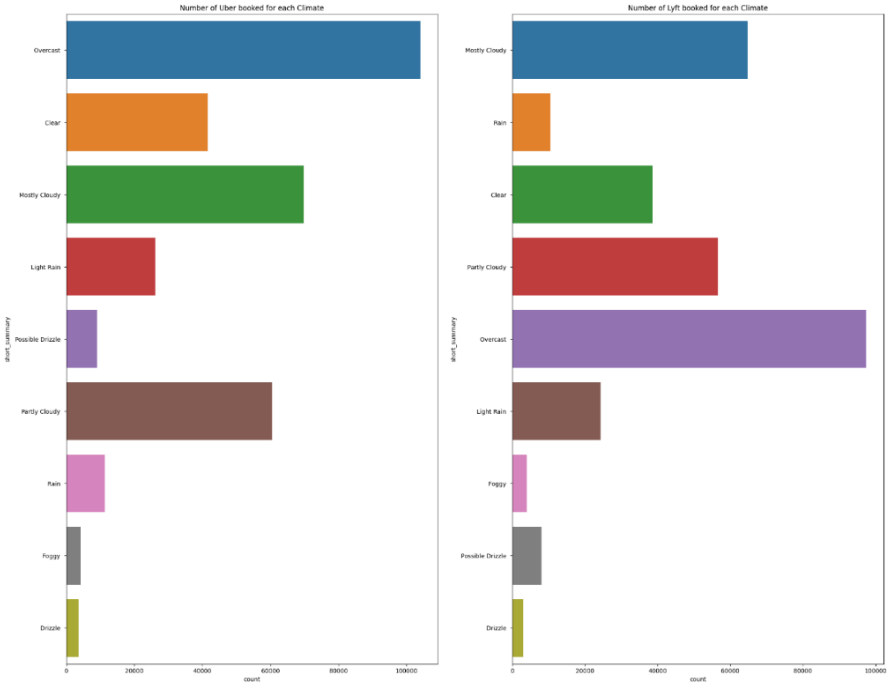



Figure 5: Number of Cabs booked in different weather conditions.

- *People tend to book cabs mostly on the overcast and the least on foggy or drizzle days in case of both Uber and Lyft.*

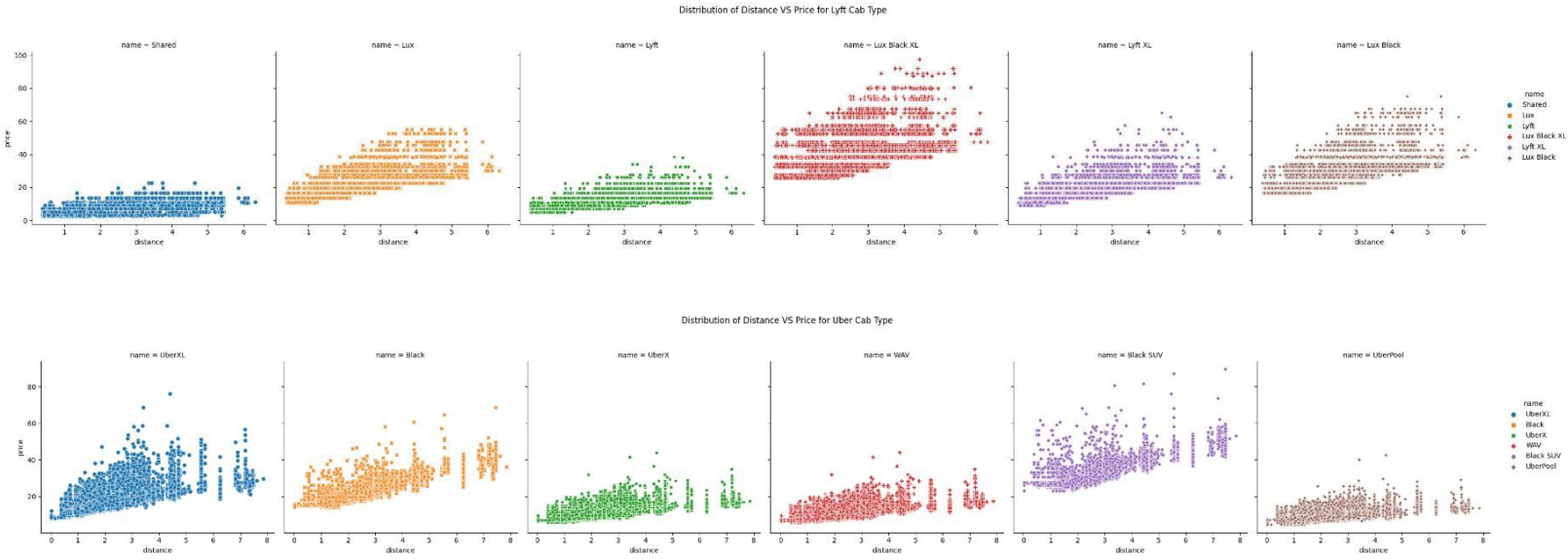



Figure 6: Scatter plot of Price vs Distance distribution

- *From the scatterplot above we see that:*

1. *In case of Uber, the UberX, WAV and UberPool are less scattered in terms of price to the distance than the other type, like Black SUV who charges more price for the same increase in the extra distance.*
2. *In case of Lyft, the Shared are less scattered in terms of price to the distance than the other type, like Lux Black XL who charges more price for the same increase in the extra distance.*

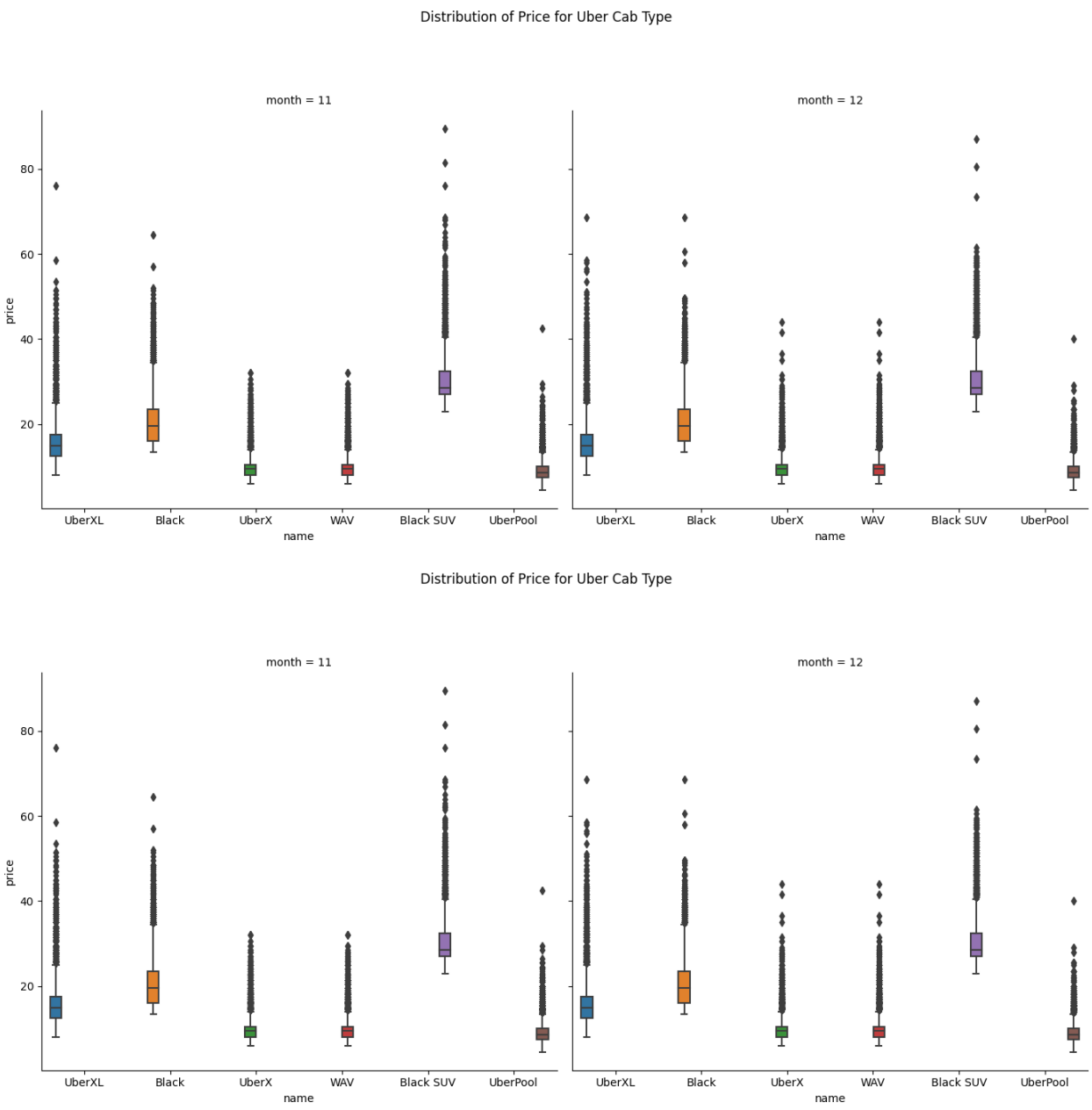



Figure 7: Boxplot plot for Price Distribution

- *The above boxplot for Uber show that:*

1. *Black SUV seems the most expensive and booking with high prices among the other type of cabs.*
2. *UberPool, WAV, UberX looks like the cheapest option as they have similar low range pattern.*

- *The above boxplot for Lyft shows that:*

1. *Lux Black XL seems the most expensive and booking with high prices among the other type of cabs.*
2. *Shared looks like the cheapest option as they have similar low range.*

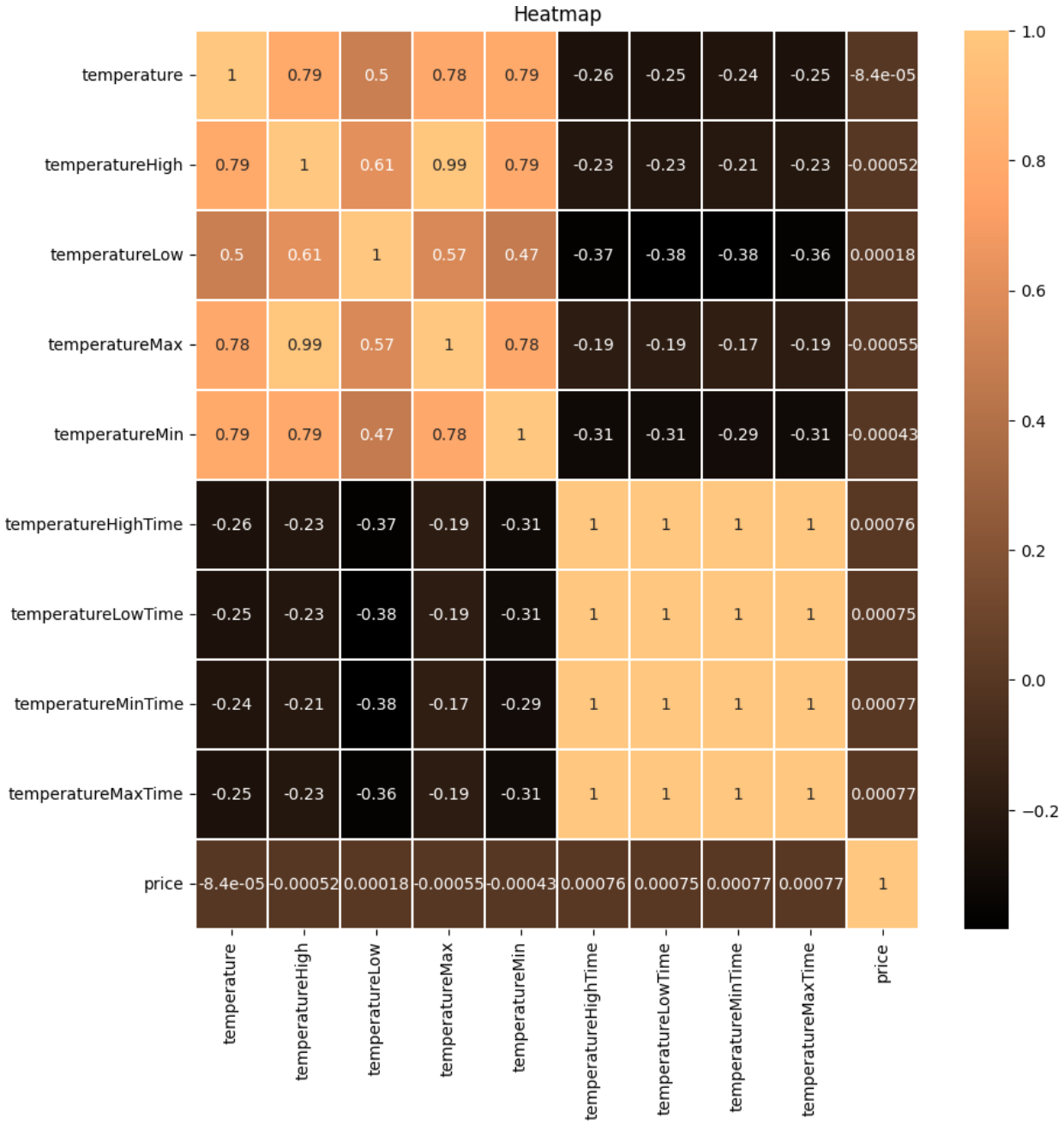



Figure 8: Heatmap

➢ *From the above correlation heatmap, we observe that the following features have no effect on the price and can thus be eliminated:*

1. *temperature*

2. *temperatureHigh*
3. *temperatureLow*
4. *temperatureMax*
5. *temperatureMin*
6. *temperatureHighTime*
7. *temperatureLowTime*
8. *temperatureMinTime*
9. *temperatureMaxTime*

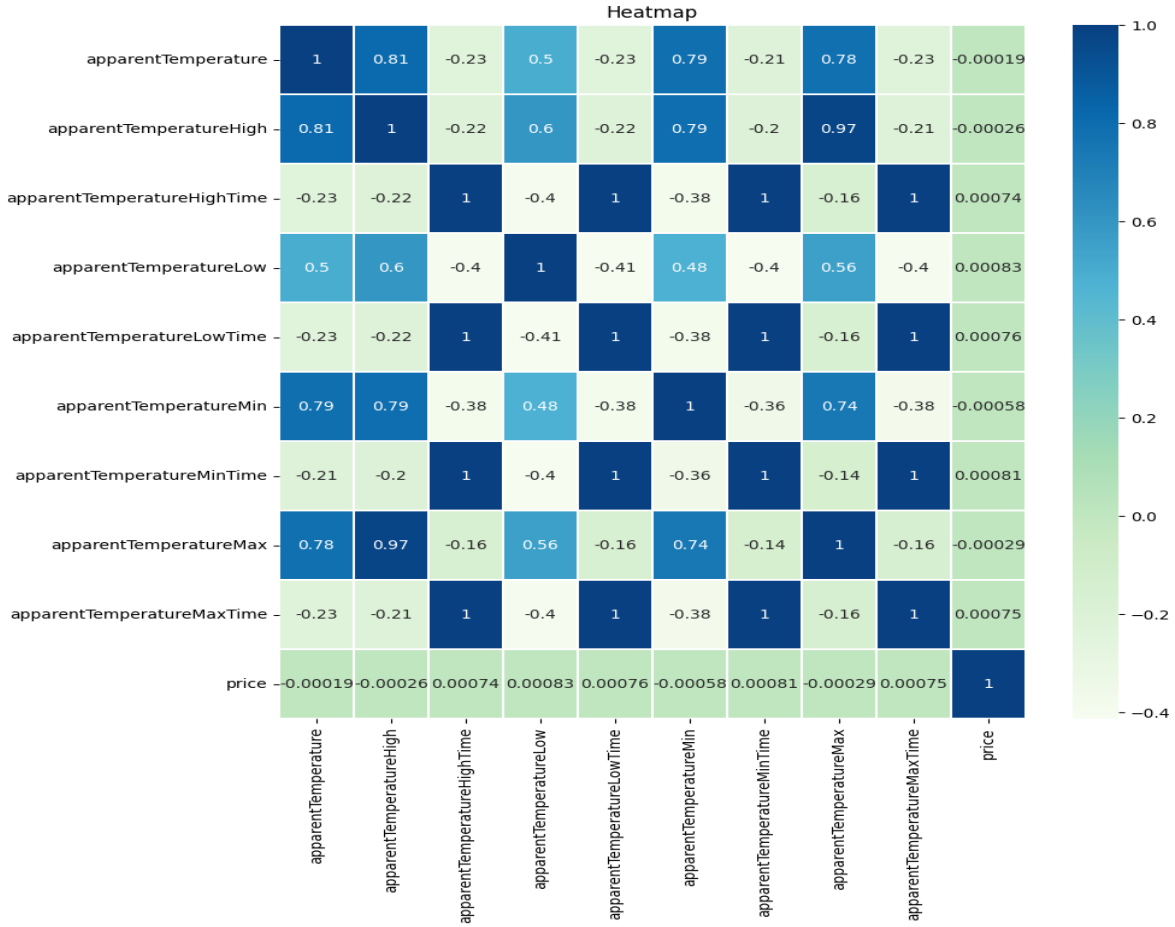



Figure 9: Heatmap

➢ *From the above correlation heatmap, we observe that the following features have no effect on the price and can thus be eliminated:*

1. *apparentTemperature*
2. *apparentTemperatureHigh*

3. *apparentTemperatureLow*
4. *apparentTemperatureMax*
5. *apparentTemperatureMin*
6. *apparentTemperatureHighTime*
7. *apparentTemperatureLowTime*
8. *apparentTemperatureMinTime*
9. *apparentTemperatureMaxTime*

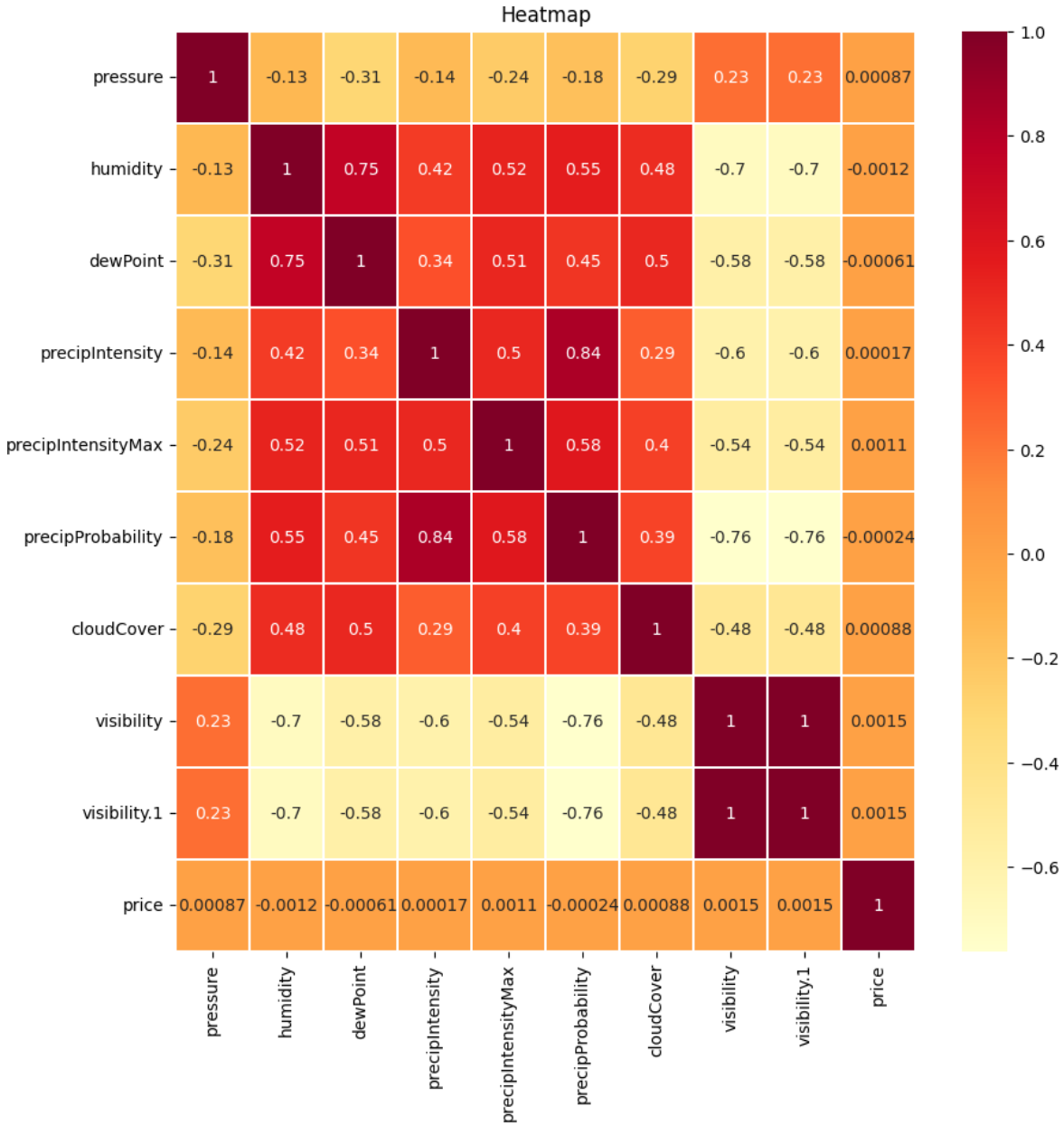



Figure 10: Heatmap

- *From the above correlation heatmap, we observe that the following features have no effect on the price and can thus be eliminated:*

1. *pressure*
2. *humidity*
3. *dewPoint*
4. *precipIntensity*
5. *precipProbability*
6. *visibility*
7. *visibility.1*
8. *cloudCover*

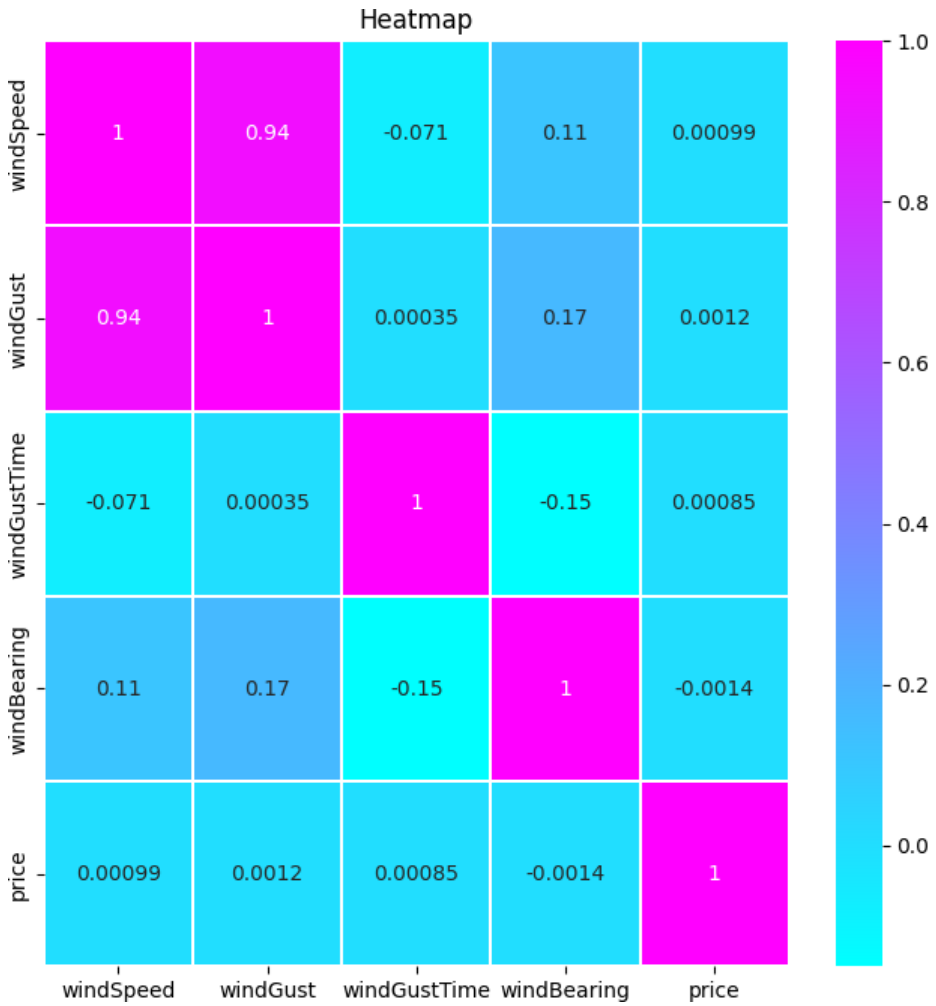



Figure 11: Heatmap

- *From the above correlation heatmap, we observe that the following features have no effect on the price and can thus be eliminated:*

1. *windSpeed*
2. *windGust*
3. *windGustTime*
4. *windBearing*

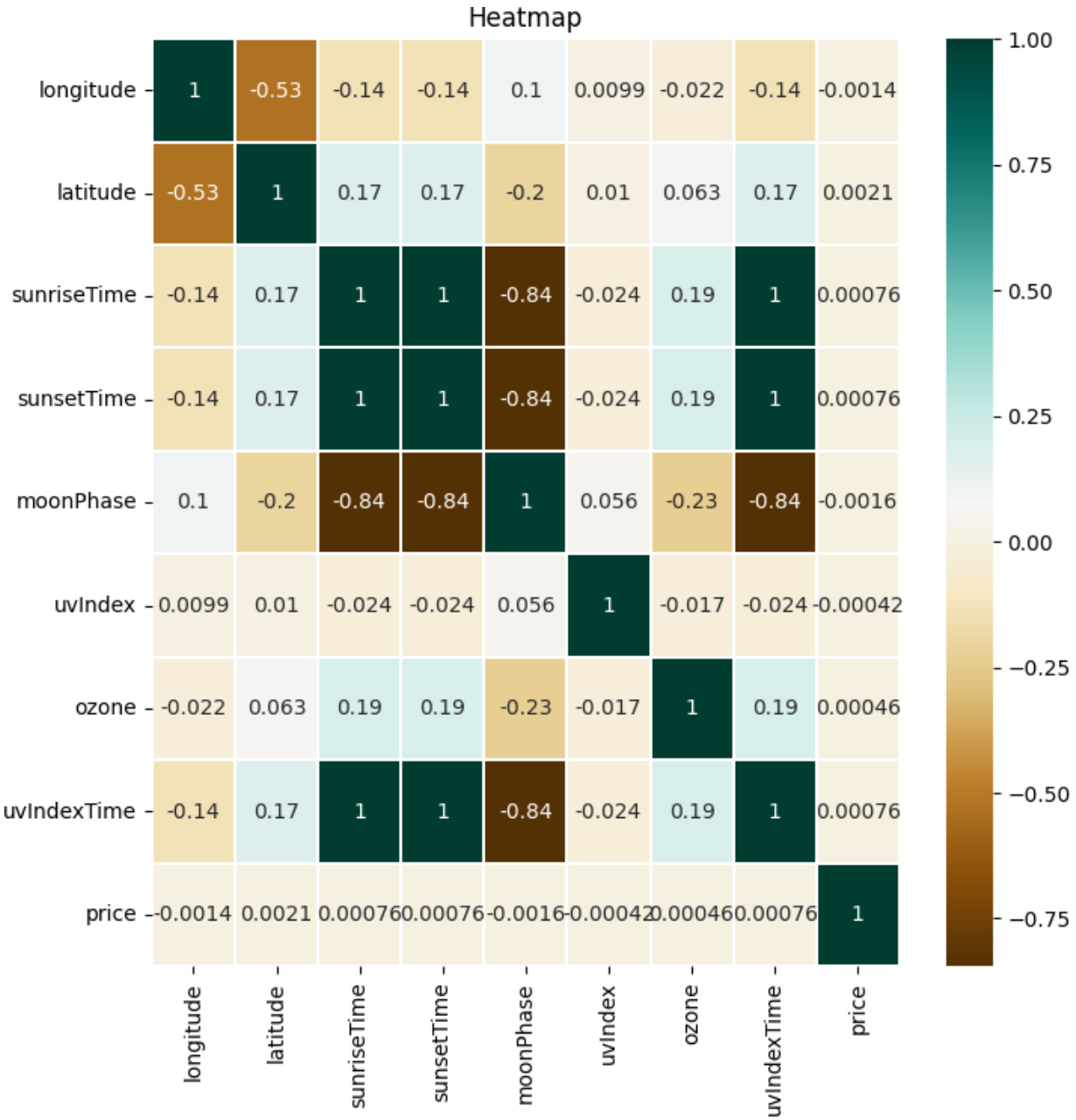



Figure 12: Heatmap

- *From the above correlation heatmap, we observe that the following features have no effect on the price and can thus be eliminated:*

1. *longitude*
2. *latitude*
3. *sunriseTime*
4. *sunsetTime*
5. *moonPhase*
6. *uvIndex*
7. *ozone*
8. *uvIndexTime*

## 6. Data Processing

- The variables that had no relation with price were removed from the model which were removed from the dataset to fit the model.
- Timestamp and id are variables are not used to fit the model, hence dropped the variables.
- Label encoding was performed on the remaining categorical variables.

## 7. Model Selection:

- After the Correlation heat map, the matrix depicted strong multi- correlation between the features.
- The regression result will help us find the Standard Error and variables with multi-collinearity.
- We will try building multiple regression models and compare their evaluation metrics to choose the best fit model for both training and validation data set.

- **Multiple Linear Regression**
  - It is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. It is used to model the relationship between a continuous response variable and continuous or categorical explanatory variable. Since our dataset contains multicollinearity, this technique could help us eliminate it.
  - **Advantages:**
    Simplicity, interpretability, and effectiveness for small datasets.
  - **Disadvantages:**
    Susceptibility to multicollinearity and outliers, and a requirement for the linearity assumption.

- **Ridge Regression Model**
  - It is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of

multicollinearity occurs, least squares are unbiased, and variances are large, this results in predicted values being far away from the actual ones.

- **Advantages:**
  Include the ability to handle multicollinearity, improved generalization performance, and a stable solution.
- **Disadvantages:**
  Include the lack of interpretability of the model and the choice of the regularization parameter.

- **Lasso Regression Model**
  - It is a type of a linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. It will help us obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso will do this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink to zero.
  - **Advantages:**
    Include feature selection, improved interpretability, and the ability to handle multicollinearity.
  - **Disadvantages:**
    Include the choice of the regularization parameter and the instability of the model when the number of predictors is greater than the number of samples.

- **Elastic Net Regression**
  - It is a type of penalized linear regression model that includes both the L1 and L2 penalties during training. Since our data dimension is greater, it could be an appropriate technique to use since it performs variable selection and regularization simultaneously.
  - **Advantages:**
    Include feature selection, improved interpretability, and the ability to handle multicollinearity.
  - **Disadvantages:**
    Include the choice of the regularization parameters and the instability of the model when the number of predictors is greater than the number of samples.

- **Polynomial Regression Method**
  - It is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. Our data set does not contain many outliers, this regression technique is suitable for predictability. Polynomial is susceptible to outliers and since we don't have many this could be a useful method.
  - **Advantages:**
    Include the ability to capture nonlinear relationships between the independent and dependent variables.
  - **Disadvantages:**
    Include the susceptibility to overfitting and the lack of interpretability of the model.

- **Decision tree Regressor**
  - A decision tree regressor is a machine learning algorithm used for regression tasks. It works by recursively splitting the dataset into smaller subsets based on the features of the data until the subsets are homogeneous with respect to the target variable.

  - In a decision tree regressor, each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a prediction of the target variable. The algorithm determines which feature to split on at each internal node based on a criterion that maximizes the information gain or minimizes the mean squared error.

  - Once the decision tree regressor is built, it can be used to predict the target variable for new data by traversing the tree from the root node to a leaf node that corresponds to the data's feature values.

  - **Advantages:**
    Decision trees are easy to understand and interpret, making them a popular choice for data analysis and decision-making. Decision trees can handle both categorical and numerical data, as well as missing values, without requiring extensive pre-processing. Decision trees can be used for both regression and classification tasks.
  - **Disadvantages:**
    Decision trees are prone to overfitting, which can lead to poor generalization performance on unseen data. Decision trees can be sensitive to small changes in the data, which can result in different tree structures. Decision trees can be biased towards features with many categories or high cardinality.

- **XGBoost Regressor**
  - In an XGBoost regressor, multiple decision trees are trained in sequence, with each subsequent tree attempting to correct the errors of the previous tree. The trees are built using a gradient descent algorithm that minimizes a loss function, such as mean squared error, and includes a regularization term that penalizes complex models.
  - **Advantages:**
    XGBoost is a state-of-the-art algorithm for supervised learning tasks that provides high accuracy and robustness. XGBoost can handle both numerical and categorical data, as well as missing values, without requiring extensive pre-processing.
  - **Disadvantages:**
    XGBoost can be computationally expensive to train, especially for large datasets with many features and samples. XGBoost requires hyperparameter tuning to obtain optimal performance, which can be time-consuming.

➢ After implementation of these regression techniques, we can compare the MAE, MSE, RMSE and R2 scores of all the regression techniques to find out which has the highest explain ability power to understand the dataset.

**8. Feature Elimination:**

- To find the optimal features and drop the redundant features, we can use either of the techniques mentioned below and compare their RMSE plots.

  - **Automated Method -RFE**
    - Recursive feature elimination is a feature selection method that fits a model and removes the weakest feature until a specified number of features is reached. It is a wrapper style feature selection algorithm that also uses filter-based feature selection internally.

    - We use Recursive Feature Elimination (RFE) to select features by recursively considering smaller and smaller sets of features.

    - First, the estimator is trained on the initial set of 56 features and the importance of each feature is obtained either through any specific attribute such as correlation co-efficient, feature importance or callable. Then, the least important features are pruned from the current set of features.

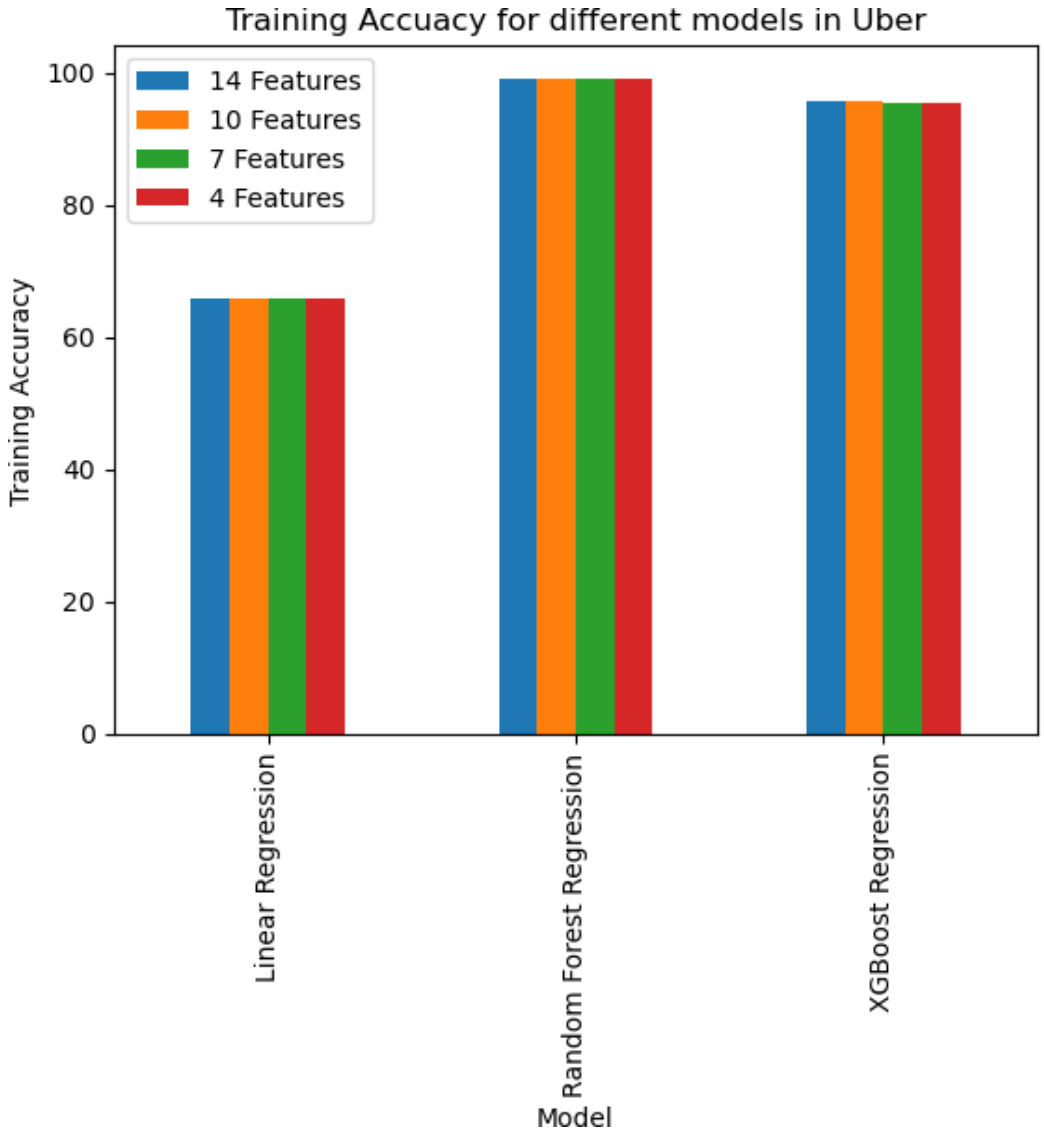
    - This procedure is recursively repeated on the pruned feature set until the desired number of features to select is eventually reached.

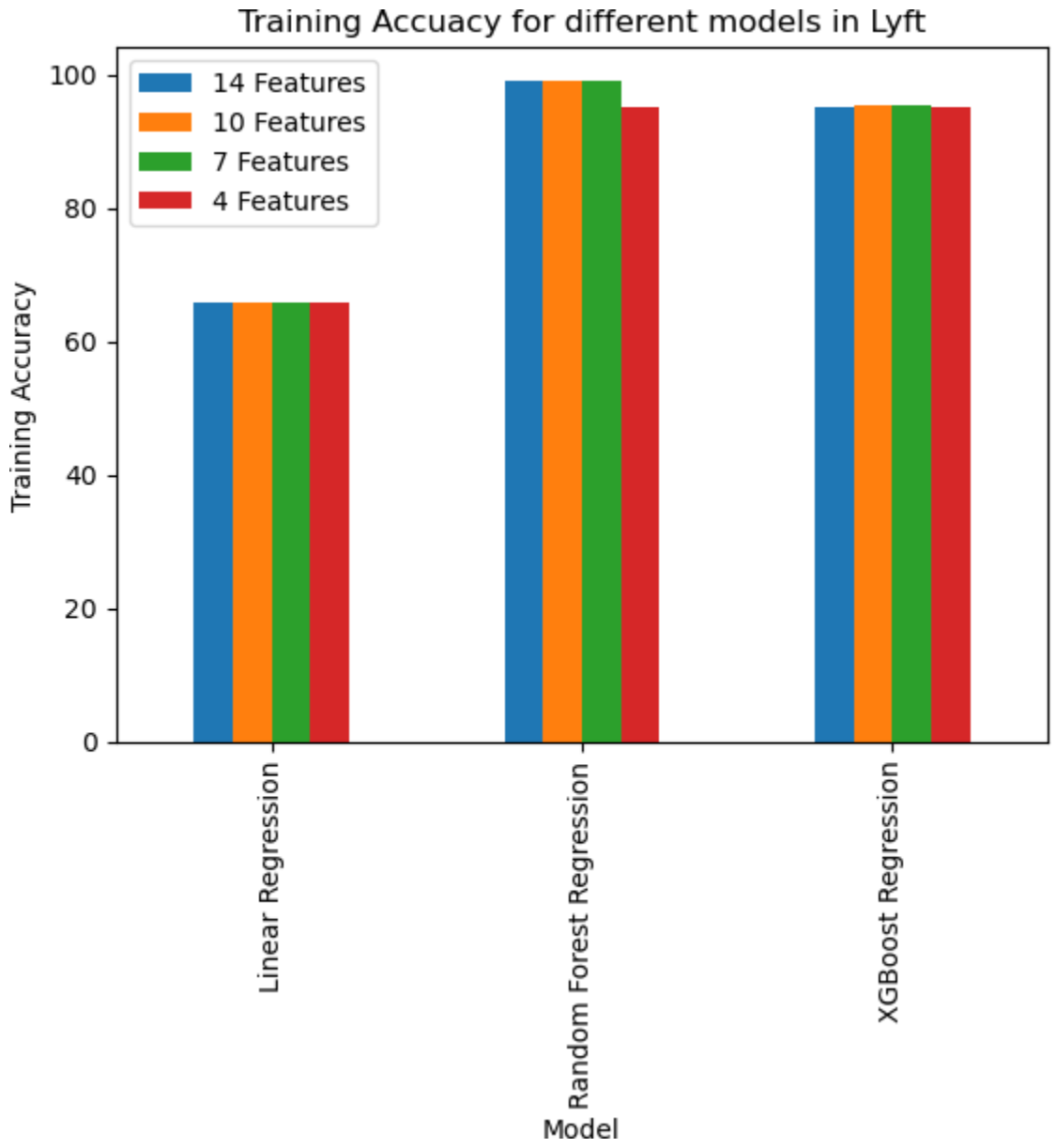
- Using the RFE method we received the following variables for each model:

| Models | Number of Features | For Uber | For Lyft |
|---|---|---|---|
| Linear Regression | 14 | 'timestamp','hour','day','month','datetime','source','destination','product_id','name','distance','surge_multiplier','short_summary','long_summary','icon' | 'timestamp','hour','day','month','datetime','source','destination','product_id','name','distance','surge_multiplier','short_summary','long_summary','icon' |
| | 10 | 'day','month','source','destination','product_id','name','distance','short_summary','long_summary','icon' | 'month','source','destination','product_id','name','distance','surge_multiplier','short_summary','long_summary','icon' |
| | 7 | 'month','source','product_id','name','distance','long_summary','icon' | 'month','source','destination','product_id','name','distance','surge_multiplier' |
| | 4 | 'source','name','distance','product_id' | 'product_id','name','distance','surge_multiplier' |
| Random Forest Regression | 14 | 'timestamp','hour','day','month','datetime','source','destination','product_id','name','distance','surge_multiplier','short_summary','long_summary','icon' | 'timestamp','hour','day','month','datetime','source','destination','product_id','name','distance','surge_multiplier','short_summary','long_summary','icon' |
| | 10 | 'timestamp','hour','datetime','destination','product_id','name','distance','surge_multiplier','short_summary' | 'timestamp','hour','day','datetime','source','destination','product_id','name','distance','surge_multiplier' |
| | 7 | 'timestamp','hour','datetime','destination','product_id','name','distance' | 'timestamp','hour','datetime','product_id','name','distance','surge_multiplier' |
| | 4 | 'datetime','product_id','name','distance' | 'product_id','name','distance','surge_multiplier' |

| | | | |
|---|---|---|---|
| XGBoost Regression | 14 | 'timestamp','hour','day','month','datetime','source','destination','product_id','name','distance','surge_multiplier','short_summary','long_summary','icon' | 'timestamp','hour','day','month','datetime','source','destination','product_id','name','distance','surge_multiplier','short_summary','long_summary','icon' |
| | 10 | 'timestamp','day','source','destination','product_id','name','distance','surge_multiplier','short_summary','icon' | 'day','source','destination','product_id','name','distance','surge_multiplier','short_summary','long_summary','icon' |
| | 7 | 'timestamp','source','destination','product_id','name','distance',’icon’ | 'source','destination','product_id','name','distance','surge_multiplier','short_summary' |
| | 4 | ‘destination','product_id','name','distance' | 'product_id','name','distance','surge_multiplier' |

Table 1: RFE Variables Selection






- On observing the training accuracy values for 14,10,7 and 4, we notice that the training accuracy is the most for 10 features in the linear regression model. Hence, we consider the 10 features.

The 10 features are as follows:

***['day','month''source','destination','product_id','name','distance','short_summary','long_summary','icon']***

## Correlation between variables:

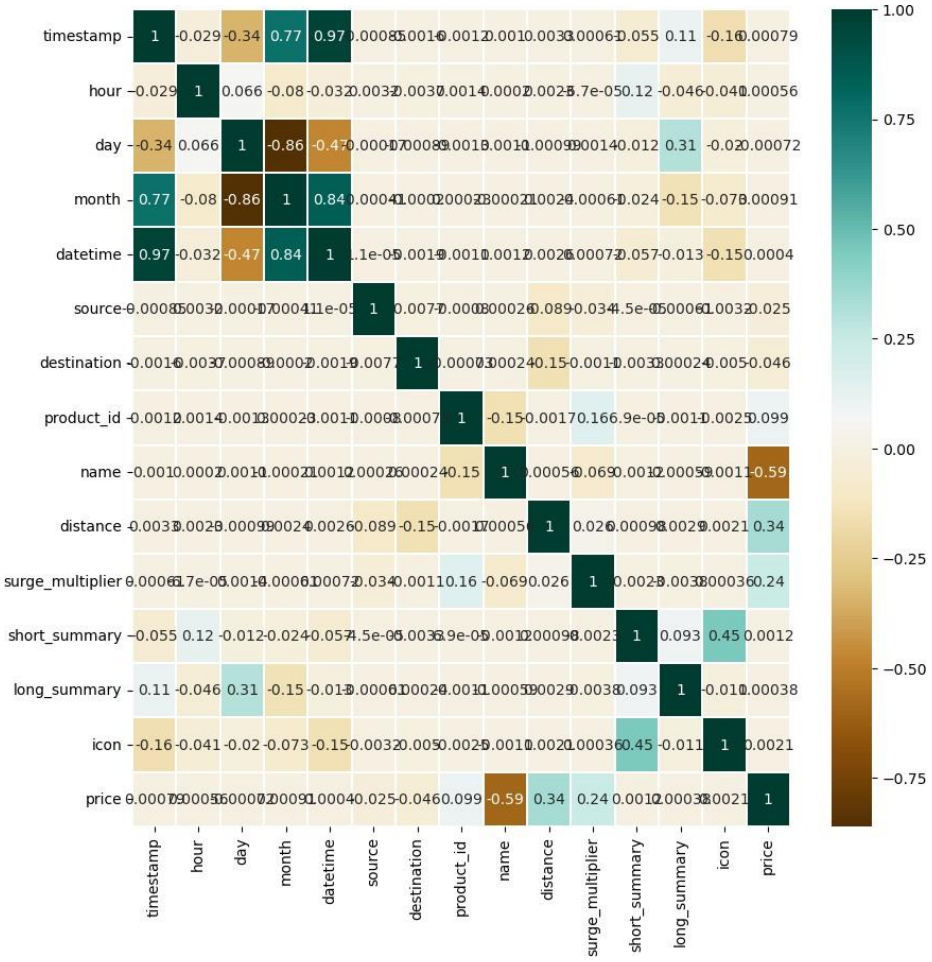



Figure 13: Heatmap for all variables for price

- We can further eliminate features in case of Uber and Lyft that have no correlation with the price as per the correlation heat map. Finally, we choose variables like :
  ***[distance, source, destination, name]***

- Distance – Distance between source and destination.
- Source – Pickup location
- Destination – Dropoff location
- Name – Type of cab booked for the ride

  (Here, source, destination and name are categorical variables and take number of the respective category as an input in the model)

## 9. Performance Evaluation:

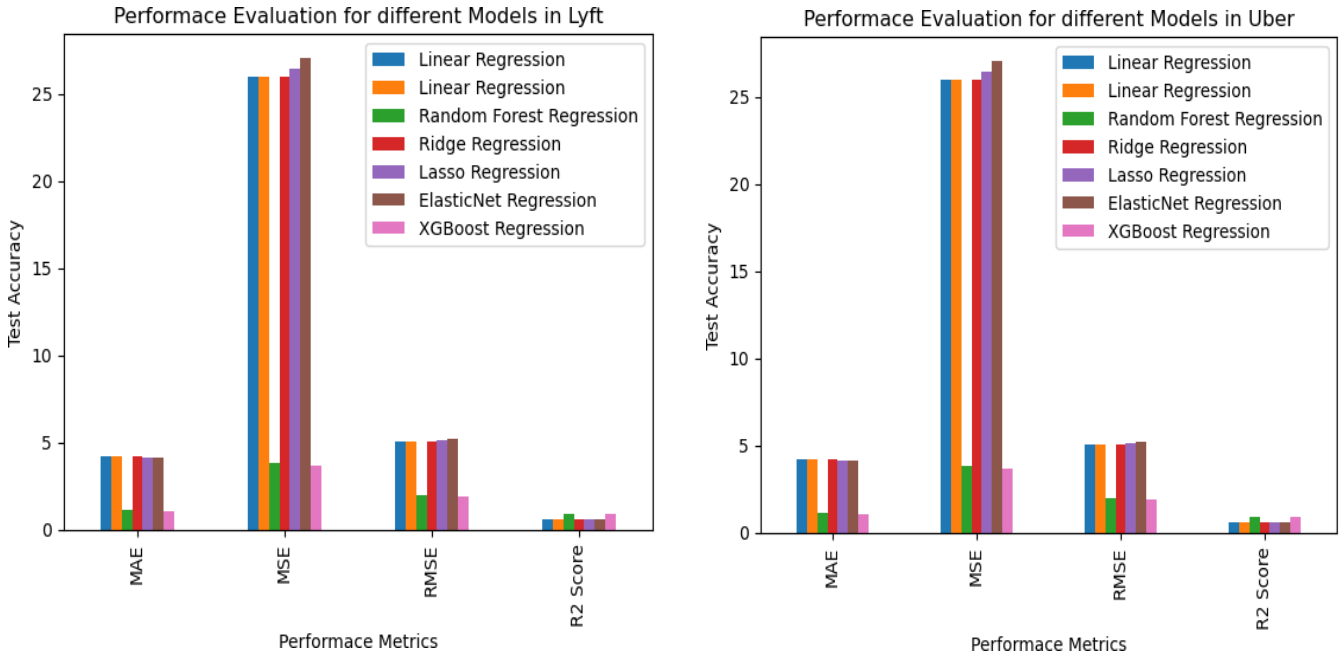



Figure 14: Performance Metrics

- Since both the XGBoostRegressor gives better results in case of Lyft and almost the same result for Uber.Hence, we select the Random Forest Regressor for the price prediction.

## 10. Impact and future scope of the project:

- The project can accommodate more scope like city, traffic conditions, type of the cab like electric or fuel and how old is the model to get more precise price. Although these factors can be included the model created is quite successful.
- Although taking all these factors into consideration it also, depends on the driving skills which cannot be quantified.